

What I really want...

- **My personal information to be accessible**
- **Published information should find me**
- **Usable anywhere**
- **Others can use what I have learned
(if I want them to)**

Wide Area Information Servers: A Supercomputer on every Desk

**Brewster Kahle
Thinking Machines Corporation**

Thinking Machines Corporation

What is it?

Electronic Publishing

(Or publishing over wires)

New Communications Technology Problems

	BOOKS	Telegraph> Telephone	Electronic Publishing
Experts only	<i>Monks</i>	<i>Operators</i>	<i>Professional searchers</i>
Distribution is hard and expensive	<i>Vellum is calf skin</i>	<i>Telephones on barb wire</i>	<i>\$1/minute over obscure modems</i>
Different interfaces	<i>1000's of languages in Europe alone</i>	<i>Switching was manual</i>	<i>//query (W5) inform?</i>
Material is intractable	<i>Scrolls and manu- scripts were about as random access as musical scores</i>	<i>No white pages</i>	<i>600 databases on Dialog ~1 Terabyte 140Gbyte at DJ 80GB card catalog at RLG</i>
Business model needed	<i>Centralized printing</i>	<i>Pay per minute</i>	<i>Not understood</i>

Navigation Techniques: Paper

- **Alphabetical Listings (dictionary, Encyclopedia)**
- **Indices (back of the book and Readers Guide)**
- **Table of Contents (outlining)**
- **Citation index**
- **"Tree of Knowledge"**
- **Have you read any good books lately?**

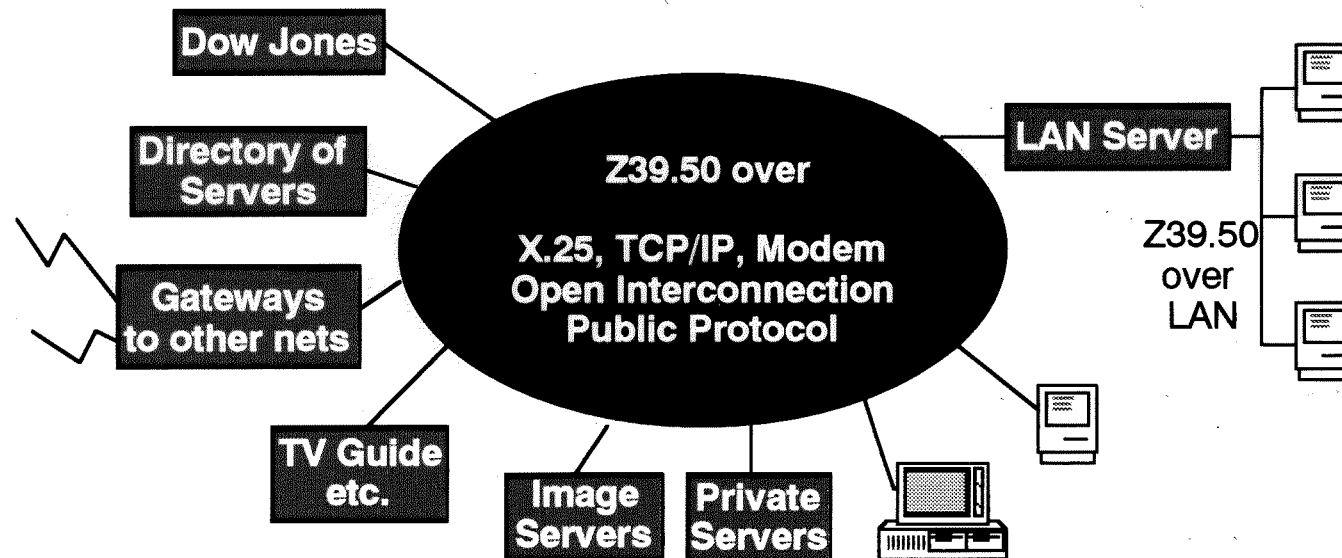
Navigation Techniques: Computers

- **Hierarchical File Systems**
- **Unix "find" and "grep", Mac "find file"**
- **Gopher, Magellan, ON Location**
- **Boolean query systems (...within 5 words of...)**
- **Static Hypertext links (see also pointers)**

Navigation Techniques: WAIS

- **English language questions and relevance feedback**
 - **Question-answer dialog**
 - **Similar to Newspapers: "More on page 5"**
 - **Dynamic Hypertext Links**
- **2 level search:**
 - **Directory of servers (server like any other)**
 - **Servers themselves**

Wide Area Information Server Architecture



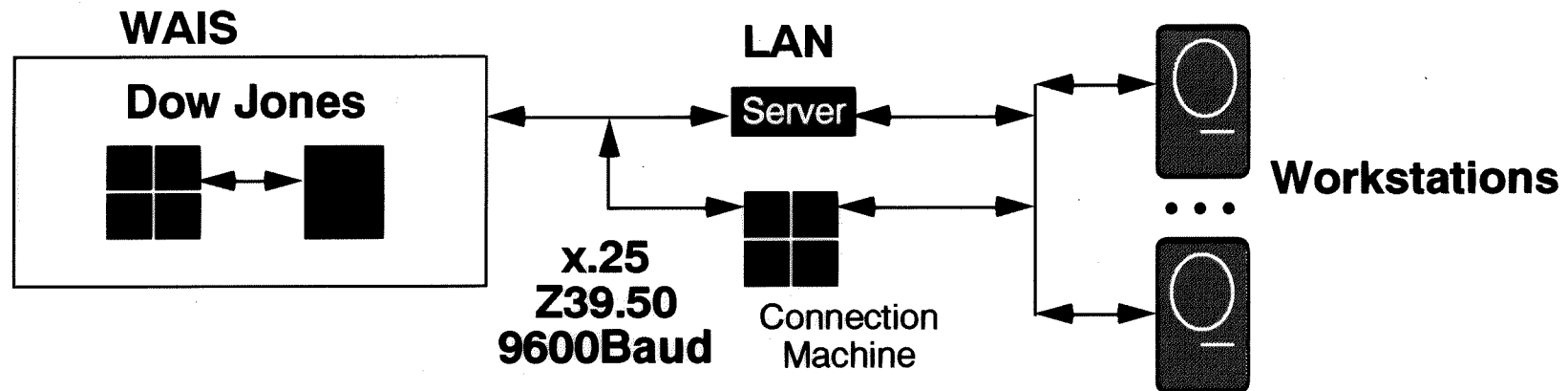
Users Needs:

- Selecting Servers
- Answering Questions
- Organizing Responses

Architecture Issues:

- Scalability
- Security
- Business model for servers
- Reliable Access

Demonstration System Structure

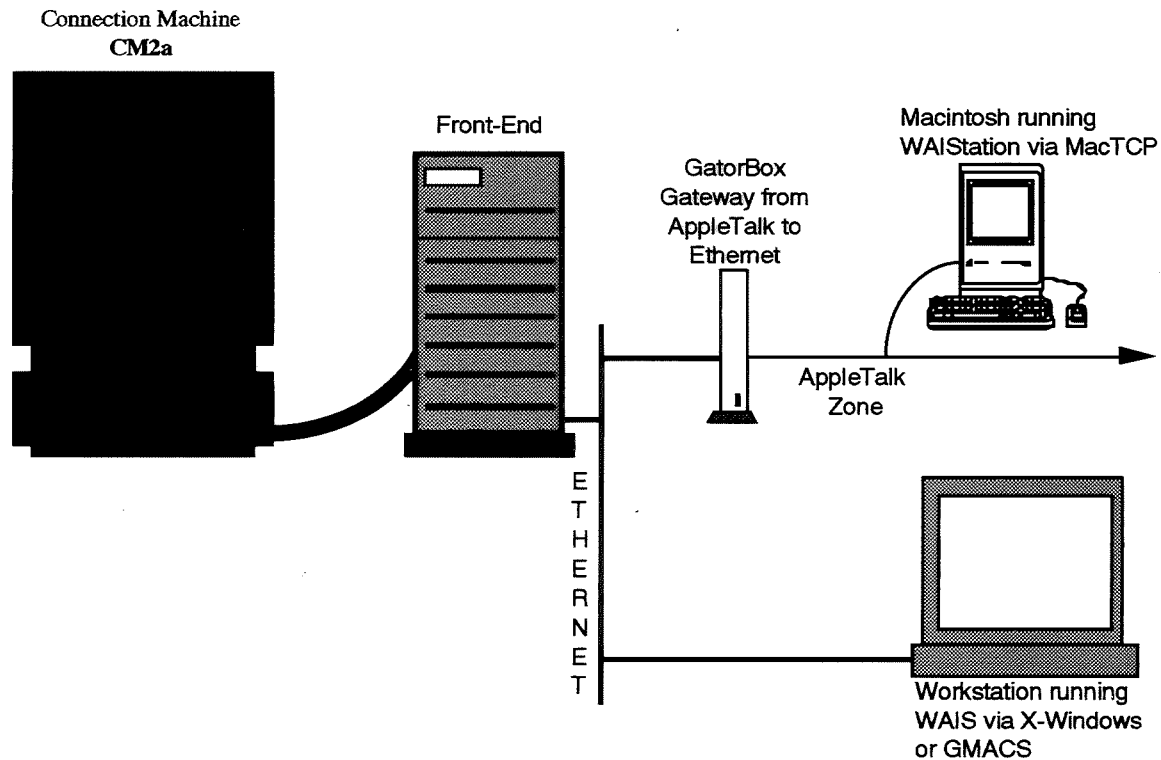


Operations:
 Archiving
 Queries
 Retrieval
IR Type:
 Broadcast
 Query by Example
Databases:
 Wall St Journal
 Barron's
 400 Business Mags

CM:
Operations:
 Queries
IR Type:
 enhanced relevance
 feedback
Databases:
 DowVision and memo's,
 mail, word processor files

Mac:
Operations:
 Human Int
 Retrieval
 Queries
 "Caching" Docs
 User Profiles
IR Type:
 Query by example
Databases:
 Personal Text
 Cached data

WAIS Hardware Components



WAIS Clients

- **Busy 24 hours a day finding information**
- **Ponder all indications of the preferences of its user**
- **Gossip with other clients about their discoveries**
- **Scours the world (within a budget) to find new sources**
- **Current implementations on PC, Macintosh, X Windows, NeXT, dumb terminal (dial-up)**

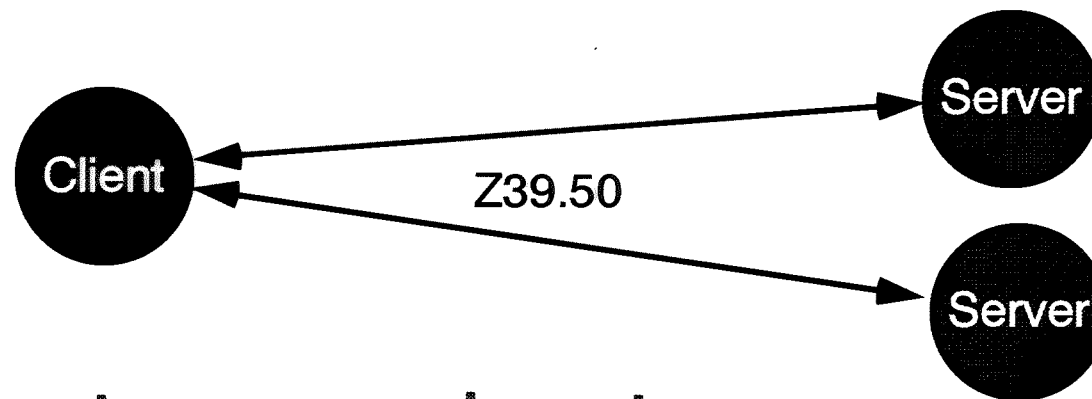
WAIS Protocol

- Based on NISO Z39.50 international standard
- Flexible — separates clients from servers
- Search: (*words, doc_ids, databases*)
returns list of: (*headline, score, doc_id, types*)
- Retrieval: (*doc_id, type, start, end*)
returns: *data of specified type*
- Doc_id: An ISBN for the Electronic Age
- Server Description Structure for the Directory of Servers

How Standard Protocol can Provides Security

- **Users do not login to server, but search only through application layer protocol (Z39.50).**
- **Server controls access to data.**
- **Network layers below application, or application layer handles authentication, encryption, billing.**

The WAIS Protocol *is* WAIS



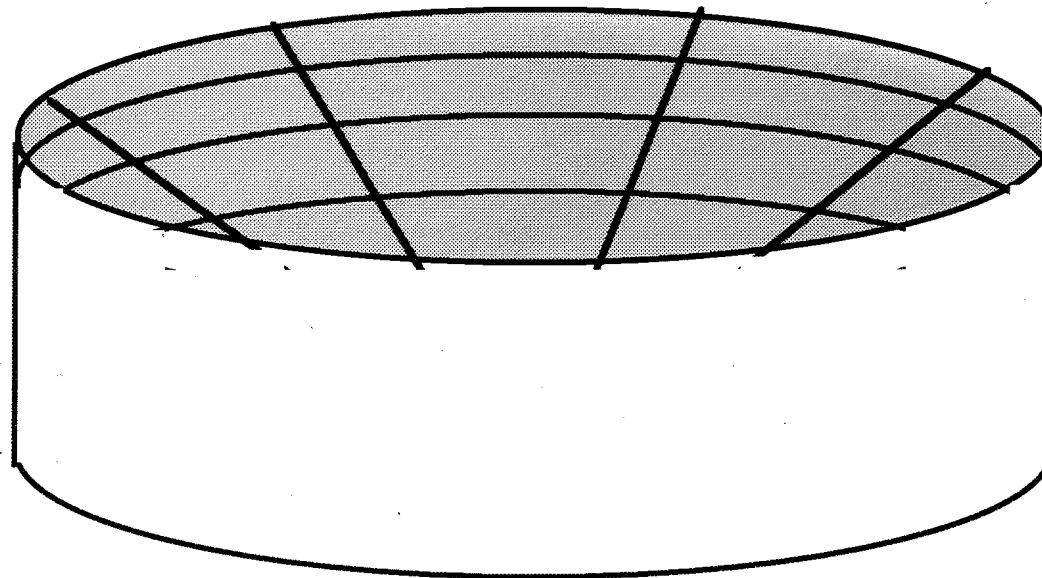
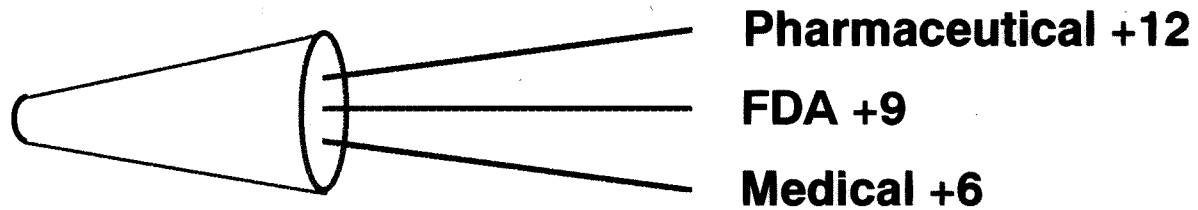
- Supports any search syntax
- Supports sophisticated clients — puts intelligence in the user's hands
- Clients can run on any platform
- Multiple servers in a single search
- Retrieve any kind of data: text, graphics, video,...

Connection Machine Server

- **1-100GBytes (and getting bigger)**
- **Supports thousands of users**
- **Automatic Indexing**
- **Uses words and phrases in question to find appropriate documents with relevance feedback, weighted term**
- **Supports Boolean Queries**
- **Cost effective hardware alternative to mainframes**

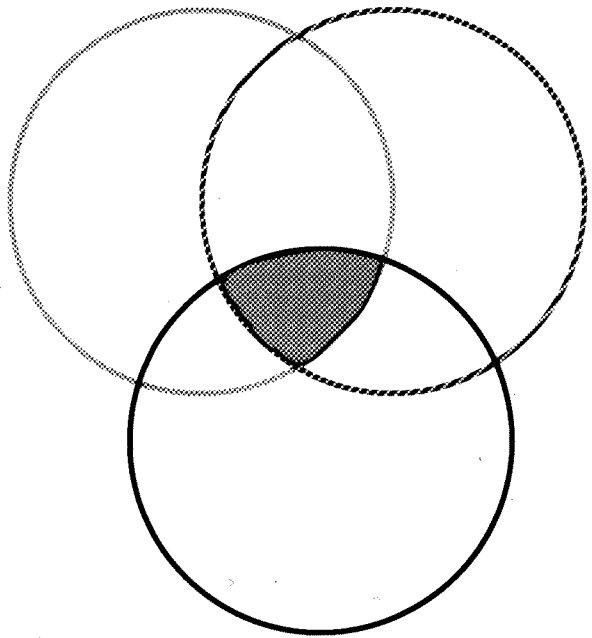
Data Parallelism:

Searching all the Documents at Once



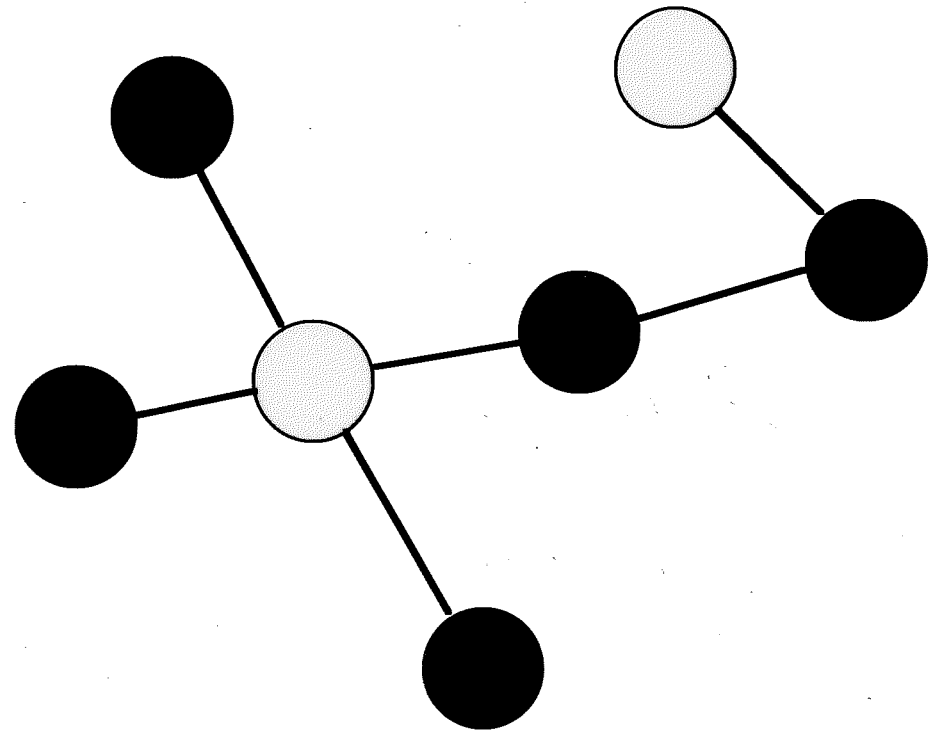
Stadium

Boolean Search



**Retrieve documents containing
specific combinations of words**

Conceptual Search



**Explore a set of documents
containing related concepts**

Boolean Query

**Hard to Use:
Complex Syntax**

**(Japanese OR Japan) AND
(building OR buildings OR (Real AND Estate) AND
(Manhattan OR (New AND York)**

**Poor Results:
The wrong information
No ranking of results**

**Have you been paying attention?...
Freer Finance: U.S. Regulators Move...
REAL ESTATE: California Initiatives...
First Boston Said To Agree on Sale Of...
Exxon, Rockefeller Group to Sell Site...
What's News--Business and Finance**

Conceptual Search: Phase 1

**Easy to Use:
No Syntax**

Japanese buying real estate in mid-town manhattan

Options:

**What do you want
to follow up?**

- 1. Time Acts to Cut Magazine Costs...**
- 2. *First Boston Said To Agree on Sale...***
- 3. Have You Been Paying Attention?**
- 4. *Exxon, Rockefeller Group to Sell Site...***
- 5. Hard Sell: Real Estate Developers...**
- 6. What's News--Business and Finance...**
- 7. Integrated Resources Buys Loft Building...**

Conceptual Search: Phase 2

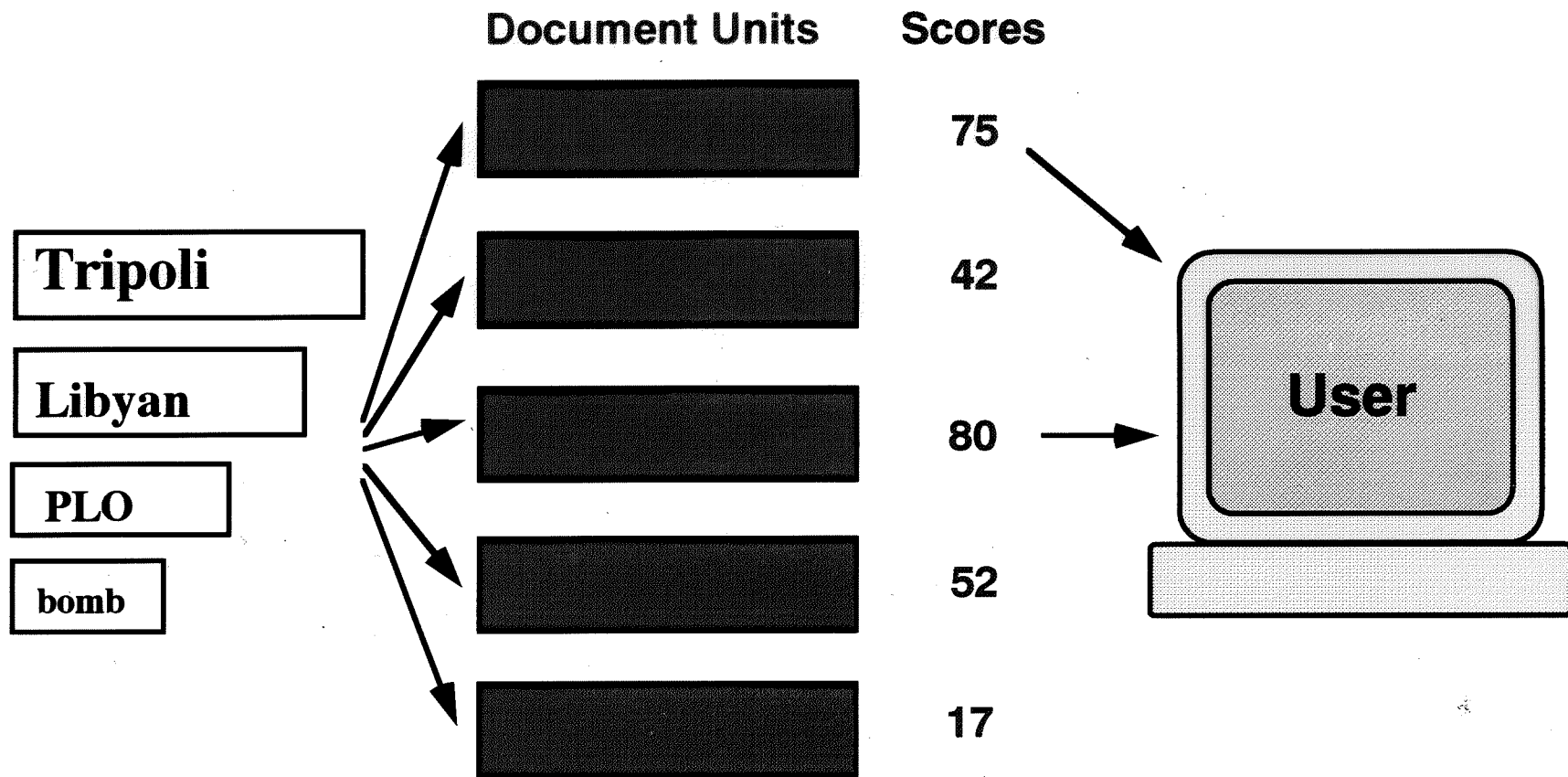
Relevance Feedback:
I like these; show
me more.

First Boston Said To Agree on Sale...
Exxon, Rockefeller Group to Sell Site...

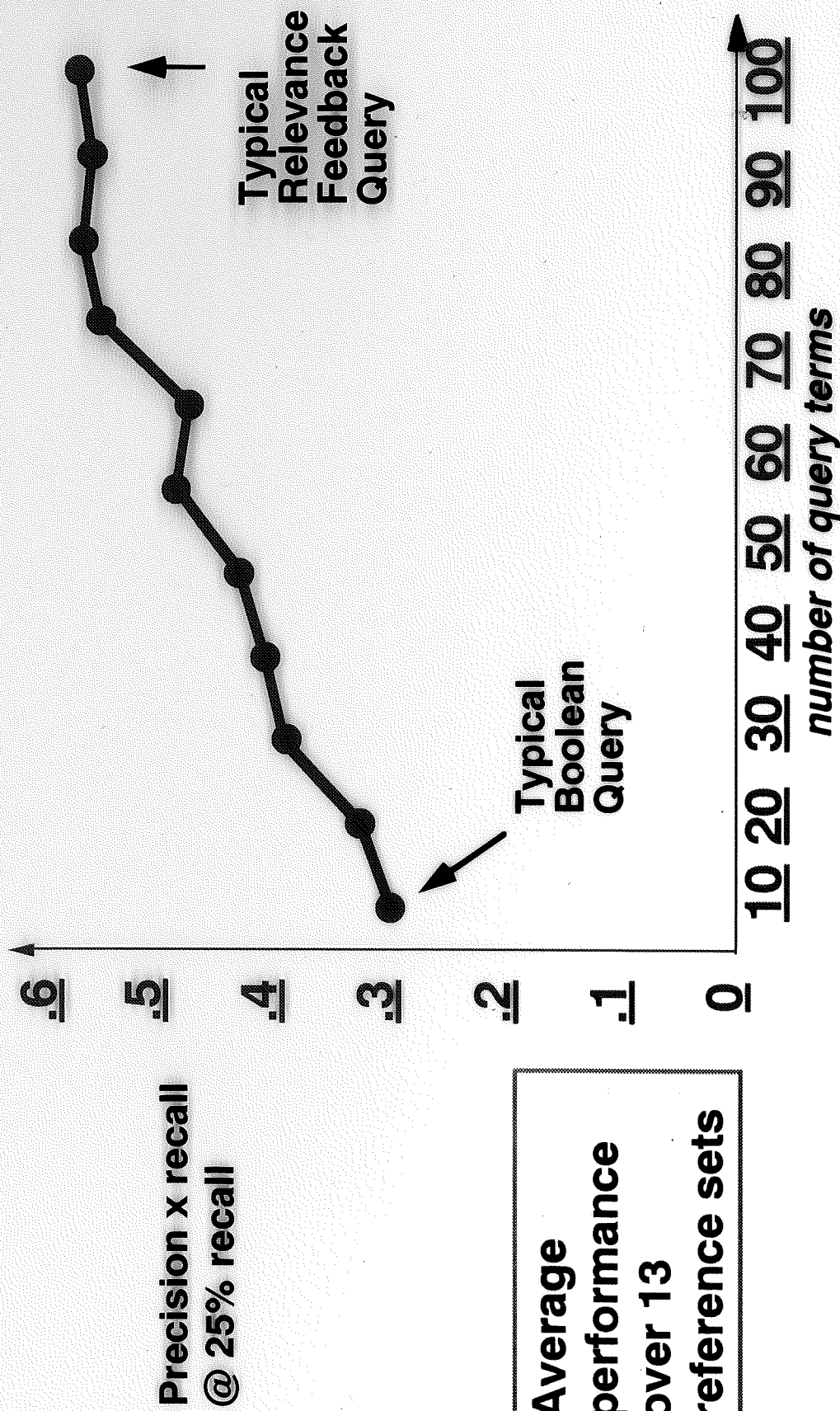
Improved results:
Articles on related
topics are found.
Results are ranked.

- 1. Bids for Exxon Building in New York...**
- 2. Time Acts to Cut Magazine Costs...**
- 3. Hard Sell: Real Estate Developers...**
- 4. Time Inc. Sells Its 45% Interest...**
- 5. Citicorp Unit Moves to Foreclose on...**
- 6. Litigious Landlords: Legal Maneuvers**

Query Broadcast To Database on Connection Machine System



Results Improve with Query Size



Document Retrieval Performance

- **Current algorithm limits:**
 - ~2 GB with 512 MB CM-2
 - ~8 GB with 2 GB CM-2
 - ~25 GB with 8 GB CM-2
 - **High recall**
 - **High precision**
 - **<< 1 sec. response**
 - **Much larger DBs searchable with CM-5 and inverted index algorithms: 100s to 1000s of Gigabytes**
- } see **Stanfill and Kahle, *Communications of the ACM*, December 1986**

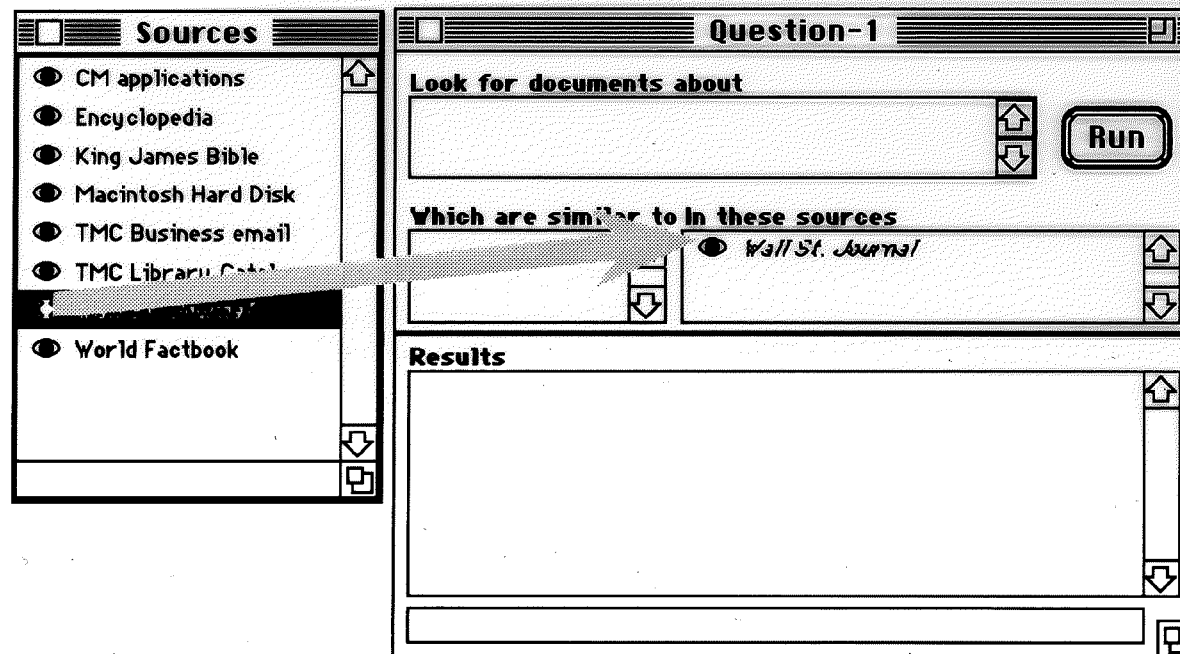


WAIStation: Active Database Sources, Saved Questions

Sources	
👁 CM applications	⬆
👁 Encyclopedia	
👁 King James Bible	
👁 Macintosh Hard Disk	
👁 TMC Business email	
👁 TMC Library Catalog	
👁 <i>Wall St. Journal</i>	
👁 World Factbook	
	⬆
	🗂

Questions	
? CM Apps Question	
? Library question	
? Encyclopedia Q	
? Patent Q	
? TMC Bus. Email Q	
? TMC Fun Q	
? Montvale Q	
? World Factbook Q	
? poetry q	
? Bible Q	

Select Data Source



Run Initial Query

Question-1

Look for documents about

recent developments in personal computers

↑

↓

Run

Which are similar to in these sources

↑

↓

👁

Wall St. Journal

↑

↓

Results

📄

*** Compaq Computer Directors Approve 2-for-1 Stock Split

↑

📄

*** International: Bull Agrees to Pay Zenith \$15 Million to End

🔍

📄

*** AT&T Set to Announce Memorex Computer Accord

🔍

📄

*** Technology Brief -- International Business Machines: Price

🔍

📄

*** Business Brief -- Data General Corp.: Four Models Are Un

🔍

📄

*** Technology: Computer Firms See the Writing on the Screen

🔍

📄

*** Retailing: Businessland Enters Japan, Aided by 4 Big Loca

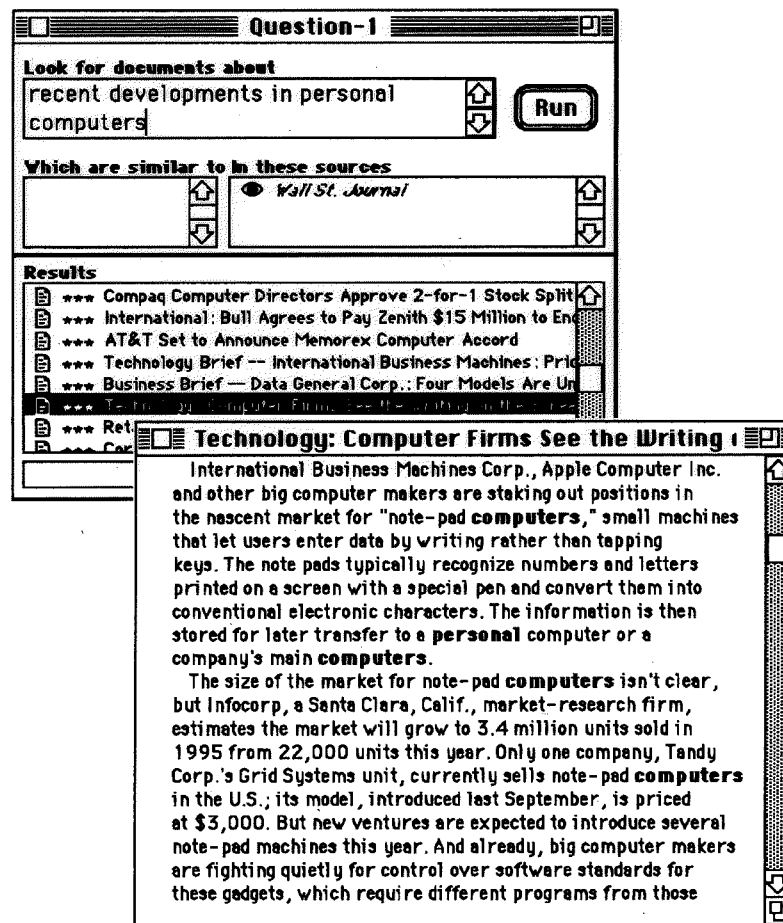
↓

📄

*** Corrections & Amplifications

🔍

Click a Headline to Display a Document



Relevance feedback: “Find me more like this one”

Question-1

Look for documents about

recent developments in personal computers

Run

Which are similar to in these sources

Technology : Cor

Wall St. Journal

Results

- *** Compaq Computer Directors Approve 2-for-1 Stock Split
- *** International: Bull Agrees to Pay Zenith \$15 Million to End
- *** AT&T Set to Announce Memorex Computer Accord
- *** Technology Brief -- International Business Machines: Price
- *** Business Brief -- Data General Corp.: Four Models Are Un
- *** Technology : Computer Firms See the Writing on the Wall
- *** Retailing: Businessland Enters Japan, Aided by 4 Big Loca
- *** Corrections & Amplifications

Relevance Feedback of Paragraph

Technology: Computer Firms See the Writing
Computer makers are scrambling to cash in on people who find the pen mightier than the keyboard.
International Business Machines Corp., Apple Computer Inc. and other big computer makers are staking out positions in the nascent market for "note-pad computers," small machines that let users enter data by writing rather than tapping keyboards. The note pads typically recognize numbers and letters printed on a screen with a special pen and convert them into computer data.

Question-1
Look for documents about
The recent developments in personal but computers
Run
Which are similar to in these sources
Technology : Cor Wall St. Journal

Results
*** Compaq Computer Directors Approve 2-for-1 Stock Split
*** International: Bull Agrees to Pay Zenith \$15 Million to End
*** AT&T Set to Announce Memorex Computer Accord
*** Technology Brief -- International Business Machines: Price
*** Business Brief -- Data General Corp.: Four Models Are Un
*** Technology: Computer Firms See the Writing on the Screen
*** Retailing: Businessland Enters Japan, Aided by 4 Big Local
*** Corrections & Amplifications

“Chaining” of Questions to Follow a Tangent

Question-1

Look for documents about
recent developments in personal computers

Which are similar to in these sources
Technology : Com Wall St. Journal

Results

- *** International: Bull Agrees to Pay Zenith \$15 Million to End
- *** AT&T Set to Announce Memorex Computer Accord
- *** Technology Brief -- International Business Machines: Pri
- *** Business Brief -- Data General Corp.: Four Models Are Un
- *** Technology : Computer Firms See the Writing on the Scree
- *** Retailing : Businessland Enters Japan, Aided by 4 Big Loca

Question-2

Look for documents about

Which are similar to in these sources
Retailing : Busin Wall St. Journal

Results

- *** Retailing : Businessland Enters Japan, Aided by 4 Big Loca
- ** What's News -- Business and Finance
- ** Technology : Computer Makers Agree on a Standard For Ne
- * Inside Track : Businessland Directors Take a Loss And Tra
- * Technology & Health : Businessland To Report Loss For 3rd
- * Technology : U.S. Computer Maker Takes on NEC on Its Ow
- * Technology : Computer Firms See the Writing on the Scree

TMC Internet Release

- **CM product for TCP/IP (complete server)**
- **Example User interfaces for free (no support)
Macintosh, Gnu Emacs, Xwindows**
- **Example unix server software to create servers**
- **Directory of Servers on the internet at least
through '92**
- **160 Servers now: Weather Maps, patents, journal
abstracts, email archives, usenet recipies,...**
- **Free Software via FTP from Think.com:/wais/***

Mailing list: wais-discussion-request@think.com

WAIS Uses

- **Over 10,000 users on the Internet**
- **Users in 24 Countries: Mexico, Singapore, Finland, Australia, etc**
- **160 Databases served from 9 Countries: Norway, Canada, UK, etc. Average 3 new databases registered per week.**

WAIS Uses: Campus Wide Info Servers

- **Class catalog and schedule**
- **Campus events: movies, sports**
- **Job listings**
- **Library catalog**
- **Phone book**
- **Professor research interests**
- **Past theses**

[<code>sol.acs.unt.edu</code>]	<code>UNTComputerDoc</code>
[<code>xantos.uio.no</code>]	<code>UiO_Publications</code>
[<code>next2.oit.unc.edu</code>]	<code>ibm.pc.FAQ</code>

WAIS Uses: Libraries

- **Easy to use card catalog**
- **Remote use from home or office**
- **Pictures, full text, scanned documents**

**[pegun.law.columbia.e]
[pegun.law.columbia.e]
[quake.think.com]**

**columbia-law-library-catalog
columbia-spanish-law-catalog
tmc-library**

WAIS Uses: Biology

- Journal Abstracts
- Sequence archives
- Images

**Currently over 20 Biology databases in
Finland, Netherlands, and US**

[<code>cmns.think.com]</code>	<code>Molecular-biology</code>
[<code>bio.vu.nl]</code>	<code>biology-compounds</code>
[<code>genbank.bio.net]</code>	<code>biology-journal-contents</code>
[<code>wais.funet.fi]</code>	<code>bionic-ai-researchers</code>
[<code>wais.funet.fi]</code>	<code>bionic-directory-of-servers</code>
[<code>wais.funet.fi]</code>	<code>bionic-enzyme</code>

WAIS Uses: Chemistry CORE Project

- **All published chemistry (8 years all ACS)**
- **Scanned pictures, ascii text**
- **Optical jukebox mass storage**
- **Connection Machine / Newton search engines**

**Project of Bellcore, ACS, Chem Abstracts,
OCLC, Cornell, and Thinking Machines**

[cujo.curtin.edu.au] [chem-eng-current-contents](#)

WAIS Uses: Business Executives

- **Dow Jones information**
- **Corporate information**
- **Personal information**

Project:
KPMG, Apple, Thinking Machines, Dow Jones

[cmns.think.com]	wall-street-journal-sample
[think.com]	Business-email

WAIS Uses: Medical Researchers/Doctors

- **Medical papers**
- **Storing and matching patient records**
- **Remote connections to specialized databases**

[wais.funet.fi] bionic-databases-limb

WAIS Uses: Community Information

- **Dial-up users: no network required**
- **Directories of services or facilities**
- **Education and entertainment**

[<code>quake.think.com]</code>	<code>internet-resource-guide</code>
[<code>sol.acs.unt.edu]</code>	<code>online-libraries</code>
[<code>quake.think.com]</code>	<code>weather</code>
[<code>lambada.oit.unc.edu]</code>	<code>nsf-bulletins</code>

Conclusion

- **Electronic Publishing can fill niches now.**
- **Companies are positioning themselves now (workstations, server, and info providers).**
- **Thinking Machines is the "Engine of the Information Industry."**